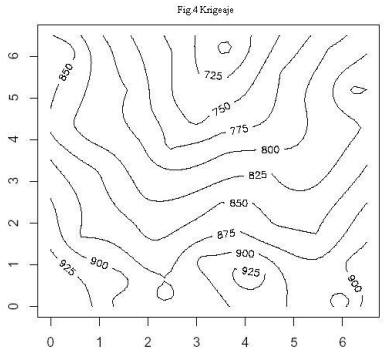
GEOESTADÍSTICA Y LOS SISTEMAS DE INFORMACIÓN GEOGRÁFICA



Ismael Díaz Alarcón
idiaz@alumnos.ubiobio.cl
Gilda Vargas MacCarte
gvargas@ubiobio.cl

Universidad del Bío Bío, Chile

Abstract
They are very diverse the occasions in which the use of the Systems of Geographical Information for to represent variables distributed spatially. Indeed new technologies facilitated graphic representations to carry out achieving results aesthetically notables. However, the most important thing is frequently neglected: the application of a best method to determine the values in those place where has not been carried out a direct measure of the variable or attribute. This work presents "geo-statistic" like a methodology to analyze correctly information coming from space samplings.

Keywords: SIG, Geo- statistic.

1. Introducción

La variabilidad espacial de algunas características de la tierra ha sido una preocupación de los investigadores prácticamente desde el principio del siglo pasado. Los principios de los métodos para datos espaciales, fueron estudiados por Halley en el año 1686, cuando trazó mapas de las direcciones del viento (Cressie, 1993). R.A. Fisher, considerado el padre de estadística 'moderna', en el año 1930, estaba consciente de la dependencia espacial, porque su teoría de diseño experimental estaba orientada a remover la influencia de esta dependencia. Sus técnicas de bloqueo y aleatorización neutralizan el efecto de correlación espacial, aunque no lo explican ni lo remueven. Krige en el año 1951 (www.Gauss.cfm.cl), trabajando con datos de concentración de oro, concluyó que con sólo la información dada por la variación sería insuficiente para explicar el fenómeno en estudio. Para lo cual, sería necesario considerar la distancia entre las observaciones. A partir de esto, surge el concepto de Geoestadística que considera la situación geográfica y la dependencia espacial. Matheron entre 1963 y 1971 (www.Monografías.com), basado en las observaciones de Krige, desarrolló la teoría de las variables regionalizadas, creando los fundamentos de la geoestadística.

La Geostadística consiste en diversos procedimientos de estimación y simulación los cuales sirven para estudiar variables distribuidas espacialmente. Esto es, a partir de un conjunto de muestras tomadas en localizaciones del dominio, en que se manifiesta un fenómeno a estudiar y consideradas representativas de su realidad, que por lo general es siempre desconocida.

Estos procedimientos permiten la descripción o caracterización de las variables con dos fines diferentes: primero, proporcionar valores estimados en localizaciones de interés y segundo, generar valores que en conjunto presenten iguales características de dispersión que los datos originales.

2. Introducción a los Métodos Geoestadísticos

2.1 Análisis de Variabilidad

El objetivo primordial se centra en estimar valores desconocidos a partir de los conocidos. Buscando el mejor estimador que minimice la varianza del error de estimación surge la Geoestadística. Se le reconoce como una rama de la estadística tradicional, que parte de la observación de que la variabilidad o continuidad espacial de las variables distribuidas en el espacio tienen una estructura particular, desarrollando herramientas matemáticas para el estudio de estas variables dependientes entre sí, llamadas variables regionalizadas. Una variable regionalizada es una función numérica con distribución espacial, que varía de un punto a otro, con continuidad clara, pero

cuyas variaciones no pueden ser representadas por una función matemática simple.

La ventaja de los modelos geoestadísticos es que reconocen variación en escala grande (tendencia) y variabilidad a escala pequeña (correlación espacial). Modelos clásicos (como superficies de respuesta, regresión no lineal) modelan variación a escala grande y suponen errores independientes. Como la forma en la que se presenta la información es muy diversa, la geoestadística se construye asumiendo condiciones de estacionariedad. Por lo que es necesario aceptar el cumplimiento de ciertas hipótesis sobre el carácter de la función aleatoria o procesos estocásticos, las cuales las describiremos a continuación.

Estacionariedad estricta: Esta primera hipótesis supone que la media de la variable aleatoria es constante en toda región, independientemente de los puntos considerados. Esta condición, como su nombre lo indica, es demasiado restrictiva en la práctica.

Estacionariedad de segundo orden: Esta hipótesis es más común en la práctica y supone que la esperanza de la variable aleatoria existe y no depende de la localización del punto, y que la función de covarianza exista y solo dependa de la distancia entre ellos.

Hipótesis Intrínseca: Estas hipótesis asumen que los incrementos son débilmente estacionarios. Al igual que las hipótesis anteriores, suponen que la esperanza matemática existe y además supone que la varianza de los incrementos es finita.

Procesos Cuasiestacionarios: Esta hipótesis es verdaderamente un compromiso de escala de la homogeneidad del fenómeno y la cantidad de datos disponibles. Esta hipótesis no es más que estacionariedad local para distancias menores que un cierto límite, donde éste representa la extensión de la región donde el fenómeno conserva cierta homogeneidad.

Estas condiciones de estacionariedad se asumen en el desarrollo teórico y en la práctica deben ser verificadas en los datos antes de comenzar un estudio geoestadístico, mediante un análisis exploratorio de datos, de modo que se refleje el grado de confiabilidad en la aplicación de estos métodos. Además, el fenómeno debe ser estacionario sólo en un cierto intervalo, por lo que es necesario comprobar si existen dichos intervalos dentro de los cuales el valor esperado y el variograma son constantes, además de contar con un número suficiente de datos que puedan contribuir a lograr una estimación precisa.

En la fase del análisis exploratorio de datos, se estudian los datos muestrales sin tener en cuenta su distribución geográfica. Es una etapa de aplicación de la estadística clásica. Se comprueba la consistencia de los datos, eliminándose aquellos que sean erróneos, y se identifican las distribuciones de las cuales provienen.

Descripción Univariable y Bivariable. En el análisis univariable podemos calcular estadísticas de resumen tales como medias, medianas, rangos, cuartiles, etc. También podemos utilizar gráficos como histogramas, tablas de frecuencia, etc. En la descripción bivariable podemos calcular medidas de relación lineal como covarianzas, correlaciones; o gráficos como diagramas de dispersión y realizar regresiones entre las variables.

Descripción Espacial: Esta parte del análisis se realiza en su mayoría en base a gráficos que nos ayudarán a reconocer la continuidad del fenómeno estudiado. Entre ellos tenemos: mapa de lugares, mapa de símbolos, mapa de contornos y gráficos en tres dimensiones.

2.2 Análisis Estructural

En la mayoría de los datos de geociencias existe continuidad espacial, es decir, datos de puntos cercanos tienen más probabilidad de ser similares que datos de puntos lejanos. Para obtener una primera impresión de la continuidad espacial, fijamos un vector h en Rd, y comparamos los datos de puntos s con los de s + h.

El h-scatterplot nos presenta información sobre la relación entre los puntos y sus vecinos en dirección h. La información tiene dos partes principales: información sobre la media, ¿cuánto más grande (o más pequeño) son los datos de los vecinos? e información sobre la covarianza, ¿qué tipo de dependencia (o correlación) tienen los puntos con sus vecinos?

La segunda parte de la información se puede evaluar con las siguientes estadísticas de resumen:

* Covariograma Muestral

$$C(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} (Z(s_i) - m(s_i))(Z(s_i + h) - m(s_i + h))$$

* Semivariograma Experimental

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (Z(s_i) - Z(s_i + h))^2$$

donde $N(h)$ es el número de pares de datos separados por una distancia h y Z es la variable aleatoria que mide el fenómeno en estudio.

El cálculo del semivarigrama experimental es la herramienta geoestadística más importante en la determinación de las características de variabilidad y correlación espacial del fenómeno estudiado, es decir, permite tener conocimiento de como la variable cambia de una localización a otra. En diversos trabajos, suelen usarse diversos algoritmos geoestadísticos sin analizar previamente las posibles estructuras espaciales, tomándose los métodos por defecto que existen en los programas utilizados. La estética de los resultados suele esconder ese gran error.

Una vez calculado los puntos del semivariograma, se representan gráficamente en función de h.

El semivariograma es una medida de desemejanza. Es decir, $\gamma(h)$ grande significa que no existe mucha relación, su confiabilidad depende mucho de la cantidad de datos involucrados en su cálculo. El gráfico de $\gamma(h)$ se caracteriza por pasar por el origen, $\gamma(0)=0$, y es en general una función creciente de h. En la mayoría de los casos, $\gamma(h)$ crece hasta cierto límite llamado meseta, en otros casos puede crecer indefinidamente. El comportamiento en el origen puede tener diferentes formas, las cuales pueden ser de forma parabólica, lineal, discontinua en el origen (efecto pepita) o discontinuo puro (ruido blanco).

2.2.1 Ajuste de un modelo teórico al semivariograma experimental

Un gráfico de semivariograma experimental, $\gamma(h)$ es formado por una serie de valores, sobre los cuales se quiere ajustar una función El modelo de semivariograma experimental calculado no será utilizado en el proceso de estimación, sino se utilizará un modelo de variograma teórico el cual debe representar fielmente la tendencia de $\gamma(h)$ en relación a h. De este modo, las estimaciones, a través del Krigeaje (estudiado más adelante), serán más exactas y, por lo tanto, más confiables.

El procedimiento de ajuste no es directo y automático, como en el caso de una regresión, por ejemplo, pero interactivo, porque en ese proceso el intérprete hace un primer ajuste y luego verifica la adaptación del modelo teórico. Dependiendo del ajuste obtenido, puede o no redefinir el modelo, hasta obtener uno que es considerado satisfactorio.

Los modelos aquí presentados son considerados modelos básicos, denominados modelos Isotrópicos. Están divididos en dos tipos de modelos: con meseta o sin meseta. Los modelos del primer tipo son referenciados por la geoestadística como modelos transitivos. Algunos de los modelos transitivos tienden a la meseta (C) asintóticamente. Para estos modelos, el alcance (α) es arbitrariamente definido como la distancia correspondiente al 95 % de la meseta. Los modelos del segundo tipo no tienen meseta, crecen a medida que la distancia aumenta. Dentro de los modelos transitivos tenemos el modelo exponencial, el esférico y el modelo Gaussiano. Dentro de los modelos sin meseta se tienen el modelo potencia y el modelo lineal

2.2.2 Análisis de anisotropía

Conviene aquí hacer un análisis sobre el comportamiento de la variabilidad del atributo en estudio. Se conoce que el semivariograma describe las características de continuidad espacial de la variable regionalizada en una dirección, pero este comportamiento puede variar según otra.

Cuando el semivariograma calculado en diferentes direcciones muestra similar comportamiento, se dice que el fenómeno es isotrópico; en caso contrario, se dice que es Anisotrópico. Los tipos de anisotropías más comunes son:

Anisotropía Geométrica. Esta presente cuando los semivariogramas en diferentes direcciones tiene la misma meseta pero distintos alcances.

Anisotropía Zonal. Esta presente cuando los semivariogramas en diferentes direcciones tiene diferentes mesetas y alcances.

2.2.3 Validación del modelo ajustado

Como el ajuste de los modelos teóricos al semivariograma experimental se realiza en forma visual o interactiva, variando los valores C_0 (efecto pepita), $C+C_0$ (meseta) y (alcance), hasta coincidir con los parámetros que mejor se ajusten, es conveniente validar el modelo seleccionado y los parámetros meseta y alcance escogidos.

La validación cruzada consiste en suprimir el i-ésimo valor medido Z (s_i) y estimarlo a partir del resto de los datos. El valor estimado de Z (s_i) se calcula por Krigeaje, procedimiento explicado más adelante. El ejemplo más común es eliminar exactamente una observación en cada paso. No podemos demostrar con la validación cruzada que el variograma es correcto, sólo comprobar que no es completamente incorrecto. Después de validar con éxito, tenemos más confianza que el variograma es más o menos insesgado y que la estimación de las varianzas del error de predicción son aproximadamente buenas.

3. Estimación: Krigeaje

El término Krigeaje se deriva del nombre Daniel G. Krige que fue el primero en introducir el uso de medias móviles para evitar la superestimación sistemática de reservas mineras. Inicialmente, el método Krigeaje fue desarrollado para resolver problemas de mapeamientos geológicos, pero su uso se extendió al mapeamiento de suelos, mapeamentos hidrológicos, atmosféricos y otros campos donde se presentaba correlación espacial.

La diferencia entre el Krigeaje y otros métodos de interpolación es la manera como los pesos son atribuidos a las diferentes .muestras. En el caso de interpolación lineal simple, por ejemplo, los pesos son todos iguales a 1/N, N es el número de muestras. En el Krigeaje, el procedimiento es similar a la interpolación por medias móviles ponderadas, sólo que aquí los pesos son determinados a partir de un análisis espacial, basado en el semivariograma experimental. Además, el Krigeaje proporciona, en promedio, estimaciones con la mínima variación.

Suponga que el objetivo es estimar el valor de Z en el punto s. El valor desconocido de Z(s) puede ser estimado a partir de una combinación lineal de n(s) valores observados, más un parámetro λ_0

$$Z^* = \lambda_0 + \sum_{i=1}^{n(s)} \lambda_i Z(s_i)$$

Se quiere un estimador insesgado, E [Z(s)-Z*(s)]=0. Esta relación impone que las medias sean iguales, por lo cual se puede demostrar que todas las versiones del Krigeaje son variaciones del predictor lineal básico $Z^* \quad m(s) \quad \lambda_0 + \sum_{i=1}^{n(s)} \lambda_i(s)\langle Z(s_i) \quad m(s_i)\rangle$

Donde *m(s)* es la media de la variable aleatoria *Z(s)*. Los pesos de ponderación $\lambda_i(s)$ y el número de observaciones usadas *n(s)* pueden cambiar de un lugar a otro.

Todas las variedades de Krigeaje tienen el objetivo de minimizar la varianza del error, con lo que es fácil demostrar que la varianza minimizada del error esta dada por

$$\sigma^2 = \sum_{i=1}^{n(s)} \sum_{j=1}^{n(s)} \lambda_i(s)\lambda_j(s) C_R(s_i - s_j) + C_R(0) - 2\sum_{i=1}^{n(s)} \lambda_i C_R(s_i - s)$$

$C_R(s_i-s_j)$, $C_R(s_i-s)$ y $C_R(0)$ los podemos calcular usando el modelo de covariograma, utilizando la relación

$\gamma_\alpha(h)=C_R(0)-C_R(h)$ (proceso residuo R(s) =Z(s)-m(s).)

El método del Krigeaje simple *(KS)* supone que la media (m) es constante en toda la región. La varianza de KS es exacta en los puntos con datos y crece cuando s se aleja de los datos. La varianza del Krigeaje no sirve como varian-

za total real porque incluye solamente el error intrínseco del método de Krigeaje. No incluye errores de medida, ni errores con respecto al ajuste del variograma o covariograma. El sistema del *KS* tiene solución única y varianza positiva si no hay datos duplicados, es decir, si ≠ sj si i ≠ j, y si el modelo del covariograma es válido.

El *Krigeaje ordinario (KO)* es el caso más común. Como también en otras variedades de Krigeaje, normalmente no se usa todos los datos para predecir Z(s), sino sólo los que están en cierta vecindad *V(s)* de s. Diferente al KS, el KO no requiere un previo conocimiento de la media *m*. Por razones numéricas, se resuelve el Sistema KO usando el covariograma en vez del variograma γ (h), es decir, se modela el variograma γ (h), pero se resuelve los sistemas de Krigeaje usando el covariograma.

Uno de los problemas encontrados al modelar semivariogramas es la tendencia en los datos, es decir, que los valores medidos aumentan o disminuyen en alguna dirección en el área en estudio. Este es el caso de un fenómeno no estacionario, lo que hace imposible la aplicación del Krigeaje presentado hasta aquí. *El Krigeaje universal* (KU) es el caso no estacionario.

4. Aplicación

Se analizaron 52 datos con coordenadas (x, y), más una variable adicional z, la cual media la altura topográfica de cierta zona. Para realizar los cálculos se utilizó el Sistema R, que es un software que contiene paquetes de estadística y programación el cual puede ser adquirido en forma gratuita en Internet. La siguiente figura muestra los datos en una malla normalizada. El tamaño del símbolo es proporcional al valor de Z.

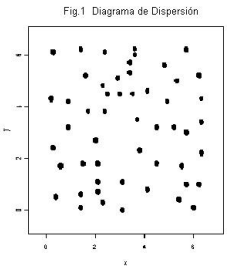


Fig. 1 Diagrama de Dispersión.

Construimos el h-scatterplot en dirección sureste-noroeste para obtener una primera impresión del comportamiento de la variable, lo que nos muestra que a medida que se aumenta la distancia entre las muestras disminuye en cierto grado la altura

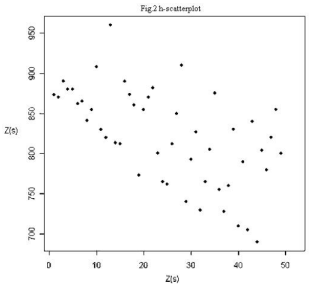


Fig.2 h-scatterplot

Luego de haber verificado la representatividad de los datos se procede a la construcción del semivariograma experimental, resultando

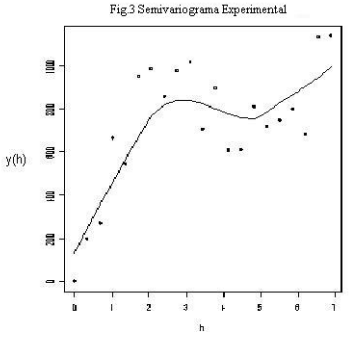


Fig.3 Semivariograma Experimental

Como se ve en el gráfico, los datos no presentan mucha relación para distancias pequeñas, pero a medida que la distancia aumenta se encuentra una mayor relación entre ellos. En otras palabras, como el fenómeno estudiado es la altura de cierto sector, se podría decir que las grandes alturas dentro de la zona se encuentran alejadas entre sí, lo cual podría tratarse de un valle rodeado por montañas.

Luego de obtenido el semivariograma experimental estimamos el fenómeno por Krigeaje, lo cual nos arroja los siguientes resultados

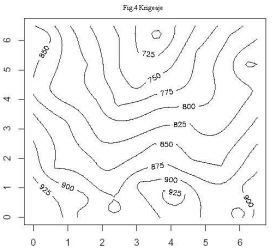


Fig.4 Krigeaje

También podemos estimar gráficamente el error de estimación, de lo cual se obtiene aproximadamente un error de 23 por sector.

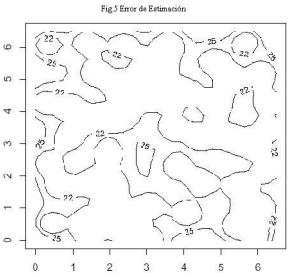


Fig.5 Error de Estimación

5. BIBLIOGRAFÍA

- www.Monografías.com
- www.Gauss.cfm.cl
- Statistics and Data Analysis in Geology; John C. Davis (1986)
- Noel A.C. Cressie: "Statistics for Spatial Data", Revised Edition, Wiley, 1993.